

Distribution-free Detection of a Submatrix

Ery Arias-Castro*

Yuchao Liu†

Abstract. We consider the problem of detecting the presence of a submatrix with larger-than-usual values in a large data matrix. This problem was considered in (Butucea and Ingster, 2013) under a one-parameter exponential family, and one of the test they analyzed is the scan test. Taking a nonparametric stance, we show that a calibration by permutation leads to the same (first-order) asymptotic performance. This is true for the two types of permutations we consider. We also study the corresponding rank-based variants and precisely quantify the loss in asymptotic power.

1 Introduction

Biclustering has emerged as an important set of tools in bioinformatics, in particular, in the analysis of gene expression data (Cheng and Church, 2000). It comes in different forms, and in fact, the various methods proposed under that umbrella may target different goals. See (Madeira and Oliveira, 2004) for a survey. Here we follow (Shabalin et al., 2009), where the problem is posed as that of discovering a submatrix of unusually large values in a (large) data matrix. For example, in the context of a microarray dataset, the data matrix is organized by genes (rows) and samples (columns). We let $\mathbf{X} = (X_{ij})$ denote the matrix, M denote the number of rows and N denote the number of columns, so the data matrix \mathbf{X} is M -by- N .

1.1 Submatrix detection

In its simplest form, there is only one submatrix to be discovered. In that context, the detection problem is that of merely detecting of the presence of an anomalous (or unusual) submatrix, which leads to a hypothesis testing problem. This was considered in (Butucea and Ingster, 2013) from a minimax perspective. Their work relies on parametric assumptions. For example, in the normal model, they assume that the X_{ij} 's are independent and normal, with mean θ_{ij} and unit variance. Under the null hypothesis, $\theta_{ij} = 0$ for all $i \in [M] := \{1, \dots, M\}$ and all $j \in [N]$. Under the alternative, there is a m -by- n submatrix indexed by $\mathcal{I}_{\text{true}} \subset [M]$ and $\mathcal{J}_{\text{true}} \subset [N]$ such that

$$\theta_{ij} \geq \theta_{\dagger}, \quad \forall (i, j) \in \mathcal{I}_{\text{true}} \times \mathcal{J}_{\text{true}}, \quad (1)$$

while $\theta_{ij} = 0$ otherwise. Here $\theta_{\dagger} > 0$ controls the signal-to-noise ratio. In that paper, Butucea and Ingster precisely establish how large θ_{\dagger} needs to be as a function of (M, N, m, n) in order for there to exist a procedure that has (worst-case) risk tending to zero in the large-sample limit (i.e., as the size of the matrix grows). They consider two tests which together are shown to be minimax optimal. One is the sum test based on

$$\text{SUM}(\mathbf{X}) = \sum_{i \in [M]} \sum_{j \in [N]} X_{ij}. \quad (2)$$

*University of California, San Diego — <http://www.math.ucsd.edu/~eariasca/>

†University of California, San Diego — <http://www.math.ucsd.edu/~yu1085/>

It is most useful when the submatrix is large. The other one is the scan test which, when the submatrix size is known (meaning m and n are known) is based on

$$\text{SCAN}(\mathbf{X}) = \max_{\mathcal{I} \subset [M], |\mathcal{I}|=m} \max_{\mathcal{J} \subset [N], |\mathcal{J}|=n} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{ij}. \quad (3)$$

When m and n are unknown, one can perform a scan test for each (m, n) in some range of interest and control for multiple testing using the Bonferroni method. From (Butucea and Ingster, 2013), and also from our own prior work, we know that the resulting procedure achieves the same first-order asymptotic performance.

To avoid making parametric assumptions, some works such as (Barry et al., 2005; Hastie et al., 2000) have suggested a calibration by permutation. We consider two somewhat stylized permutation approaches:

- *Unidimensional permutation.* The entries are permuted within their row. (One could permute within columns, which is the same after transposition.)
- *Bidimensional permutation.* The matrix is vectorized, the entries are permuted uniformly at random as one would in a vector, and the matrix is reformed.

The first method is most relevant when one is not willing to assume that the entries in different rows are comparable. It is appealing in the context of microarray data and was suggested, for example, in (Hastie et al., 2000). The second method is most relevant in a setting where all the variables are comparable. In the parlance of hypothesis testing, the first method derives from a model where the entries within each row are exchangeable under the null, while the second method arises when assuming that all the entries are exchangeable under the null.

Contribution 1 (Calibration by permutation). We analyze the performance of the scan test when calibrated using one of these two permutation approaches. We show that, regardless of the variant, the resulting test is (first order) asymptotically as powerful as a calibration by Monte Carlo with full knowledge of the parametric model. We prove this under some standard parametric models.

Remark 1. We focus on the scan statistic (3) and abandon the sum statistic (2) for at least two reasons: 1) the sum statistic cannot be calibrated without knowledge of the null distribution; 2) the sum statistic is able to surpass the scan statistic when it is impossible to locate the submatrix with any reasonable accuracy, which is somewhat less interesting to the practitioner.

A calibration by permutation is computationally intensive in that it requires the repeated computation of the test statistic on permuted data. In practice, several hundred permutations are used, which can cause the method to be rather time-consuming. A possible way to avoid this is to use ranks, which was traditionally important before the availability of computers with enough computational power. (Hettmansperger, 1984) is a classical reference. In line with the two permutation methods described above, we consider the corresponding methods for ranking the entries:

- *Unidimensional ranks.* The entries are ranked relative to the other entries in their row.
- *Bidimensional ranks.* The entries are ranked relative to the all other entries.

The use of ranks has the benefit of only requiring calibration (typically done on a computer nowadays) once for each matrix size $M \times N$. It has the added benefit of yielding a method that is much more robust to outliers.

Contribution 2 (Rank-based method). We analyze the performance of the scan test when the entries are replaced by their ranks following one of the two methods just described. We show that, regardless of the variant, there is a mild loss of asymptotic power, which we precisely quantify. We do this under some standard parametric models.

1.2 More related work

The scan statistic (3) is computationally intractable and there has been efforts to offer alternative approaches. We already mentioned (Shabalin et al., 2009), which proposes an alternate optimization strategy: given a set of rows, optimize over the set of columns, and vice versa, alternating in this fashion until convergence to a local maximum. This is the algorithm we use in our simulations. It does not come with theoretical guarantees (other than converging to a local maximum) but performs well numerically. A spectral method is proposed in (Cai et al., 2015) and a semidefinite relaxation is proposed in (Chen and Xu, 2014). These methods can be run in time polynomial in the problem size (meaning in M and N). (Ma and Wu, 2015) establishes a lower bound based on the Planted Clique Problem that strictly separates the performance of methods that run in polynomial time from the performance of the scan statistic.

Our work here is not on the computational complexity of the problem. Rather we assume that we can compute the scan statistic and proceed to study it. In effect, we contribute here to a long line of work that studies permutation and rank-based methods for nonparametric inference. Most notably, we continue our recent work (Arias-Castro et al., 2015) where we study the detection problem under a similar premise but under much more stringent structural assumptions. The setting there would correspond to an instance where the submatrix is in fact a block, meaning, that $\mathcal{I}_{\text{true}}$ and $\mathcal{J}_{\text{true}}$ are of the form $\mathcal{I}_{\text{true}} = \{i + 1, \dots, i + k\}$ and $\mathcal{J}_{\text{true}} = \{j + 1, \dots, j + l\}$. The present setting assumes much less structure. The related applications are very different in the end. Nevertheless, the technical arguments developed there apply here with only minor adaptation. The main differences are that we consider two types of permutation and ranking protocols.

1.3 Content

The rest of the paper is organized as follows. In Section 2 we describe a parametric setting where likelihood methods have been shown to perform well. This parametric setting will serve as benchmark for the nonparametric methods that ensue. In Section 3 we consider the detection problem and study the scan statistic with each of the two types of calibration by permutation. In Section 4 we consider the same problem and study the rank-based scan statistic using each of the two types of rankings. In Section 5 we present some numerical experiments on simulated data. All the proofs are in Section 6.

2 The parametric scan

Following the classical line in the literature on nonparametric tests, we will evaluate the nonparametric methods introduced later on a family on parametric models. As in (Butucea and Ingster, 2013), and in our preceding work (Arias-Castro et al., 2015), we consider a one-parameter exponential family in natural form.

To define such a family, fix a probability distribution ν on the real line with zero mean and unit variance, and with a sub-exponential right tail, specifically meaning that $\varphi(\theta) := \int_{\mathbb{R}} e^{\theta x} \nu(dx) < \infty$ for some $\theta > 0$. Let θ_* denote the supremum of all such $\theta > 0$. (Note that θ_* may be infinite.) The

family is then parameterized by $\theta \in [0, \theta_*)$ and has density with respect to ν defined as

$$f_\theta(x) = \exp\{\theta x - \log \varphi(\theta)\}. \quad (4)$$

By varying ν , we obtain the normal (location) family, the Poisson family (translated to have zero mean), and the Rademacher family.

Such a parametric model is attractive as a benchmark because it includes these popular models and also because likelihood methods are known to be asymptotically optimal under such a model. Butucea and Ingster (2013) showed this to be the case for the problem of detection, where the generalized likelihood ratio test is based on the scan statistic (3).

Under such a parametric model, the detection problem is formalized as a hypothesis testing problem where ν plays the role of null distribution. In detail, suppose that the submatrix is known to be $m \times n$. The search space is therefore

$$\mathbb{S}_{m,n} := \{\mathcal{S} = \mathcal{I} \times \mathcal{J} : \mathcal{I} \subset [M], |\mathcal{I}| = m \text{ and } \mathcal{J} \subset [N], |\mathcal{J}| = n\}. \quad (5)$$

We assume that the X_{ij} 's are independent with $X_{ij} \sim f_{\theta_{ij}}$, and the testing problem is

$$H_0 : \theta_{ij} = 0, \quad \forall (i, j) \in [M] \times [N], \quad (6)$$

versus

$$H_1 : \exists \mathcal{S}_{\text{true}} \in \mathbb{S}_{m,n} \text{ such that } \begin{cases} \theta_{ij} \geq \theta_{\dagger}, & \forall (i, j) \in \mathcal{S}_{\text{true}}, \\ \theta_{ij} = 0, & \text{otherwise.} \end{cases} \quad (7)$$

Here θ_{\dagger} controls the signal-to-noise ratio is assumed to be known in this formulation.

In this context, we have the following.

Theorem 1 (Butucea and Ingster (2013)). *Consider an exponential model as described above, with ν having finite fourth moment. Assume that*

$$M, N, m, n \rightarrow \infty, \quad \frac{m}{M}, \frac{n}{N} \rightarrow 0, \quad \frac{\log(M \vee N)}{m \wedge n} \rightarrow 0. \quad (8)$$

Then the sum test based on (2), at any fixed level $\alpha > 0$, has limiting power 1 when

$$\theta_{\dagger} \frac{mn}{\sqrt{MN}} \rightarrow \infty. \quad (9)$$

Then the scan test based on (3), at any fixed level $\alpha > 0$, has limiting power 1 when

$$\liminf \frac{\theta_{\dagger} \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} > 1. \quad (10)$$

Conversely, the following matching lower bound holds. Assume in addition that $\log M \asymp \log N$ and $m \asymp n$. Then any test at any fixed level $\alpha > 0$ has limiting power at most α when

$$\theta_{\dagger} \frac{mn}{\sqrt{MN}} \rightarrow 0 \quad \text{and} \quad \liminf \frac{\theta_{\dagger} \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} < 1. \quad (11)$$

We note that Butucea and Ingster (2013) derived their lower bound under slightly weaker assumptions on M, N, m, n .

Remark 2. Proper calibration in this context is based on knowledge of the null distribution ν . In more detail, consider a test that rejects for large values of a statistic $T(\mathbf{X})$. Assuming a desired level of $\alpha > 0$ and that ν is either diffuse or discrete (for simplicity), the critical value for T is set at t_α , where $t_\alpha = \inf\{t : \nu(T(\mathbf{X}) \geq t) \leq \alpha\}$. The test is then $\mathbb{I}\{T(\mathbf{X}) \geq t_\alpha\}$. In practice, t_α may be approximated by Monte Carlo sampling.

3 Permutation scan tests

In the previous section we described the work of [Butucea and Ingster \(2013\)](#), who in certain parametric models show that the sum test (2) and scan test (3) are jointly optimal for the problem of detecting a submatrix. This is so if they are both calibrated with full knowledge of the null distribution (denoted ν earlier).

What if the null distribution is unknown? A proven approach is via permutation. This is shown to be optimal in some classical settings ([Lehmann and Romano, 2005](#)) and was recently shown to also be optimal in more structured detection settings ([Arias-Castro et al., 2015](#)). We prove that this is also the case in the present setting of detecting a submatrix. We consider the two types of permutation, unidimensional and bidimensional, described in Section 1.1. More elaborate permutation schemes have been suggested, e.g., in ([Barry et al., 2005](#)), but these are not considered here, in part to keep the exposition simple. Indeed, we simply aim at showing that a calibration by permutation performs very well in the present context.

Let Π be a subgroup of permutations of $[M] \times [N]$, identified with $[MN]$. Then a calibration by permutation of the scan statistic (or any other statistic) yields the P-value

$$\mathfrak{P}(\mathbf{X}) = \frac{\#\{\pi \in \Pi : \text{SCAN}(\mathbf{X}_\pi) \geq \text{SCAN}(\mathbf{X})\}}{|\Pi|}, \quad (12)$$

where $\mathbf{X}_\pi = (X_{\pi(i,j)})$ is the matrix permuted by π . The permutation scan test at level α is the test $\mathbb{I}\{\mathfrak{P}(\mathbf{X}) \leq \alpha\}$. It is well-known that this is a valid P-value, in the sense that, under the null, it dominates the uniform distribution on $[0, 1]$ ([Lehmann and Romano, 2005](#)). (This remains true of a Monte Carlo approximation.)

The set of unidimensional permutations, denoted Π_1 , is that of all permutations that permute within each row, while the set of bidimensional permutations, denoted Π_2 , is simply the set of all permutations. Obviously, $\Pi_1 \subset \Pi_2$ with $|\Pi_1| = (N!)^M$ and $|\Pi_2| = (MN)!$, and they are both groups.¹

Theorem 2. *Consider an exponential model as described in Section 2. In addition to (8), assume*

$$\log^3(M \vee N)/(m \wedge n) \rightarrow 0, \quad (13)$$

and that either (i) ν has support bounded from above, or (ii) $\max_{i,j} \theta_{ij} \leq \bar{\theta}$ for some $\bar{\theta} < \theta_$ fixed. Let the group of permutations Π be either Π_1 or Π_2 ; if $\Pi = \Pi_1$, we require that $\varphi(\theta) < \infty$ for some $\theta < 0$. Then the permutation scan test based on (12), at any fixed level $\alpha > 0$, has limiting power 1 when (10) holds.*

The additional condition (on ν or the nonzero θ_{ij} 's) seems artificial, but just as in ([Arias-Castro et al., 2015](#)), we are not able to eliminate it. Other than that, in view of Theorem 1 we see that the permutation scan test — just like the parametric scan test — is optimal to first-order under a general one-parameter exponential model.

4 Rank-based scan tests

Rank tests are classical special cases of permutation tests ([Hettmansperger, 1984](#)). Traditionally, when computers were not as readily available and not as powerful, permutation tests were not practical, but rank tests could still be, as long as calibration had been done once for the same (or

¹The group structure is important. See the detailed discussion in ([Hemerik and Goeman, 2014](#)).

a comparable) problem size. Another well-known advantage of rank tests is their robustness to outliers.

We consider the two ranking protocols described in Section 1.1. After the observations are ranked, the distribution under the null is the permutation distribution, either uni- or bi-dimensional depending on the ranking protocol. This is strictly true under an appropriate exchangeability condition, which holds in the null model we consider here where all observations are IID. In fact, the unidimensional rank scan test is a form of unidimensional permutation test, and the bidimensional rank scan is a form of bidimensional permutation test, each time, the statistic being the rank scan

$$\text{SCAN}(\mathbf{R}) = \max_{\mathcal{I} \subset [M], |\mathcal{I}|=m} \max_{\mathcal{J} \subset [N], |\mathcal{J}|=n} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} R_{ij}, \quad (14)$$

where $\mathbf{R} = (R_{ij})$ is the matrix of ranks.

Rank tests have been studied in minute detail in the classical setting (Hájek and Sidak, 1967; Hettmansperger, 1984). Typically, this is done, again, by comparing their performance with the likelihood ratio test in the context of some parametric model. Typically, there is some loss in efficiency, unless one tailors the procedure to a particular parametric family.² Such a performance analysis was recently carried out for the rank scan in more structured settings (Arias-Castro et al., 2015). We again extend this work here and obtain the following.

Define

$$\Upsilon = \mathbb{E}(Z \mathbb{1}_{(Z>Y)}) + \frac{1}{2} \mathbb{E}(Z \mathbb{1}_{(Z=Y)})$$

where Y, Z are IID with distribution ν . (This is the same constant introduced by Arias-Castro et al. (2015).)

Theorem 3. *Consider an exponential model as described in Section 2. Assume that (8) holds. Let the group of permutations Π be either Π_1 or Π_2 . The rank scan test at any fixed level $\alpha > 0$ has limiting power 1 when*

$$\liminf \frac{\theta_{\dagger} \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} > \frac{1}{2\sqrt{3}\Upsilon}. \quad (15)$$

The proof is omitted as it is entirely based on an adaptation of that of Theorem 2 and arguments given in (Arias-Castro et al., 2015) to handle the rank moments.

Compared to the (optimal) performance of the parametric and permutation scan tests in the same setting (Theorem 1 and Theorem 2), we see that there is a loss in power. However, the loss can be quite small. For example, as argued in (Arias-Castro et al., 2015), in the normal model $1/(2\sqrt{3}\Upsilon) \leq \sqrt{\pi/3} \approx 1.023$.

5 Numerical experiments

We performed some numerical experiments³ to assess the accuracy of our asymptotic theory. To do so, we had to deal with two major issues in terms of computational complexity. The first issue is the computation of the scan statistic defined in (3). There are no known computationally tractable method for doing so. As Butucea and Ingster (2013) did, we opted instead for an approximation in the form of the alternate optimization (or hill-climbing) algorithm of Shabalin et al. (2009). Since in principle this algorithm only converges to a local maximum, we run the algorithm on several

²Actually, Hajek (1962) proposes a more complex method that avoids the need for knowing the null distribution.

³In the spirit of reproducible research, our code is publicly available at <https://github.com/nozoeli/NPDetect>

random initializations and take the largest output. The second issue is that of computing the permutation P-value defined in (12). (This is true for the permutation test and also for the special case of the rank test.) Indeed, examining all possible permutations in Π (either Π_1 or Π_2) is only feasible for very small matrices. As usual, we opted for Monte Carlo sampling. Specifically, we picked π_1, \dots, π_B IID uniform from Π with $B = 500$ in our setup. We then estimate the permutation P-value by

$$\hat{\mathfrak{P}}(\mathbf{X}) = \frac{\#\{b \in [B] : \text{SCAN}(\mathbf{X}_{\pi_b}) \geq \text{SCAN}(\mathbf{X})\}}{B + 1}. \quad (16)$$

We mention that when rank methods are applied, the ties in the data are broken randomly.

Simulation setup Our simulation strategy is as follows. A data matrix \mathbf{X} of size $M \times N$ is generated with the anomaly as $[m] \times [n]$. All the entries of \mathbf{X} are independent with distribution f_0 (same as ν) except for the anomalous ones which have distribution $f_{\theta_{\dagger}}$ for some $\theta_{\dagger} > 0$. We compare the permutation tests and rank tests (unidimensional and bidimensional) with the scan test calibrated by Monte Carlo (using 500 samples), which serves as an oracle benchmark as it has full knowledge of the null distribution f_0 . By construction, all tests have the prescribed level. As we increase θ_{\dagger} , the P-values of the different tests are recorded. Each setting is repeated 200 times.

As one of the main purposes of our simulations is to confirm our theory, we zoom in on the region near the critical value

$$\theta_{\text{crit}} = \sqrt{\frac{2(m \log \frac{M}{m} + n \log \frac{N}{n})}{mn}}, \quad (17)$$

which comes from (10). Specifically, we increase θ_{\dagger} from $0.5 \times \theta_{\text{crit}}$ to $1.5 \times \theta_{\text{crit}}$ with step size $0.125 \times \theta_{\text{crit}}$ to explore the behavior of P-values around the critical value.

The Normal Case Here we generate data from normal family, where f_{θ} is $\mathcal{N}(\theta, 1)$. We used two setups, $(M, N, m, n) = (200, 100, 10, 15)$ and $(M, N, m, n) = (200, 100, 30, 10)$, to assess the performance of the tests under different anomaly sizes. The resulting boxplots of the averaged P-values are shown in Figure 1.

From the plots we see that the P-values are generally very close to 0 when θ_{\dagger} exceeds θ_{crit} . When $(m, n) = (10, 15)$ the convergence towards 0 is slower, which may be due to the small size of the anomalous submatrix. As expected, the (oracle) Monte Carlo test is best, followed by the bidimensional permutation test, followed by the unidimensional permutation test. That said, the differences appear to be minor, which confirms our theoretical findings.

For the rank tests, we observe a similar behavior of the P-values, with the bidimensional showing superiority over the unidimensional rank test, but the loss of power with respect to the oracle test is a bit more substantial, as predicted by the theory. As shown before, $1/(2\sqrt{3}\Upsilon) \approx 1.03$ for the standard normal, so that we should place the critical threshold approximately at $1.03 \times \theta_{\text{crit}}$. This appears to be confirmed in the setting where $(m, n) = (30, 10)$. While the P-values for the rank tests converge relatively slowly when $(m, n) = (10, 15)$ (for unidimensional rank test the P-value is close to 0 at $\theta = 1.5 \times \theta_{\text{crit}}$), this may be due to the relatively small size of the anomaly.

The Poisson Case As another example, we consider the Poisson family, where f_{θ} corresponds to $\text{Poisson}(e^{\theta}) - 1$. The data matrix and anomaly sizes are the same as they are in the normal case. The resulting boxplots of the P-values are shown in Figure 2. Overall, we observe a similar behavior of the P-values.

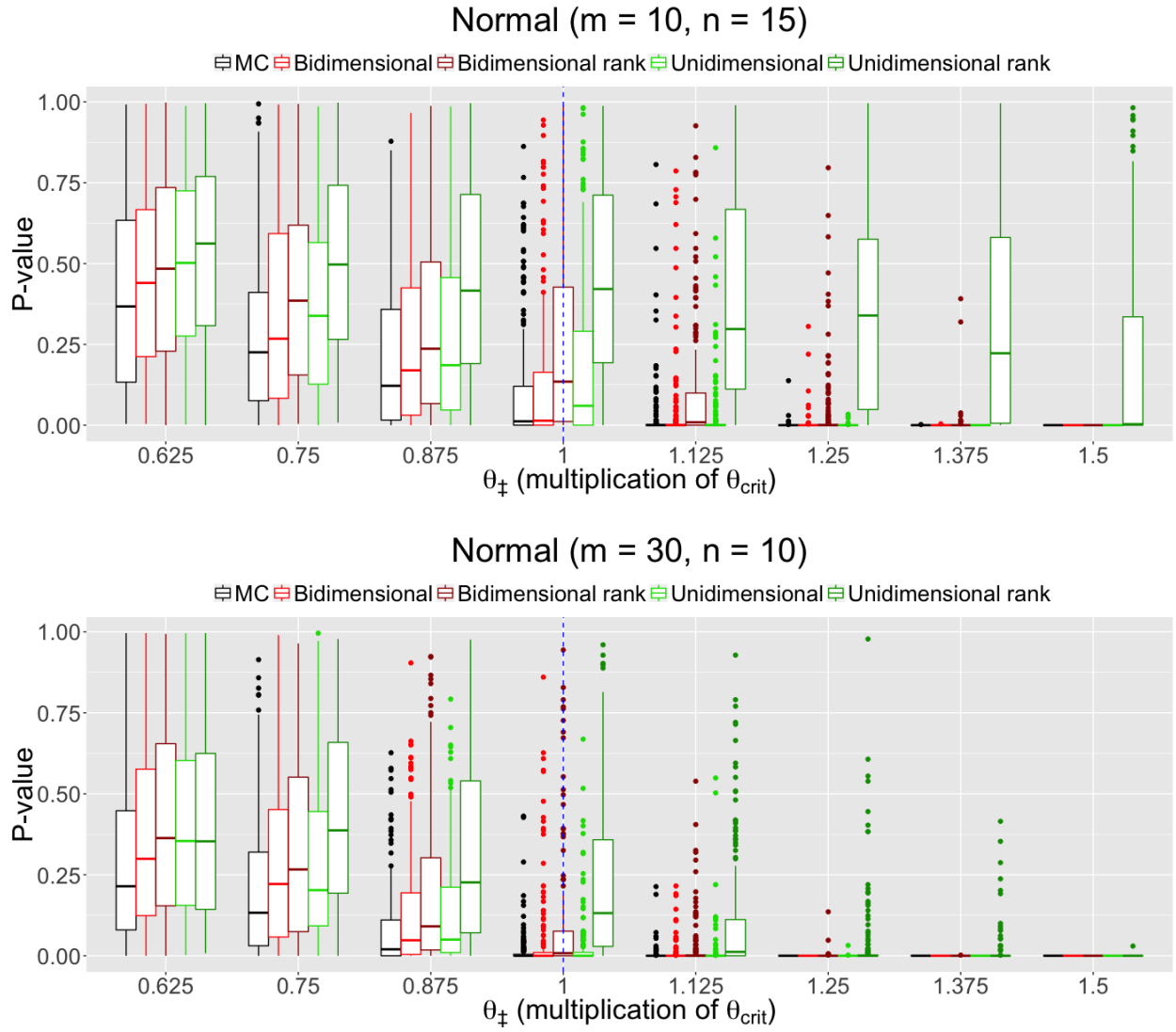


Figure 1: P-values of various forms of scan tests in the normal model

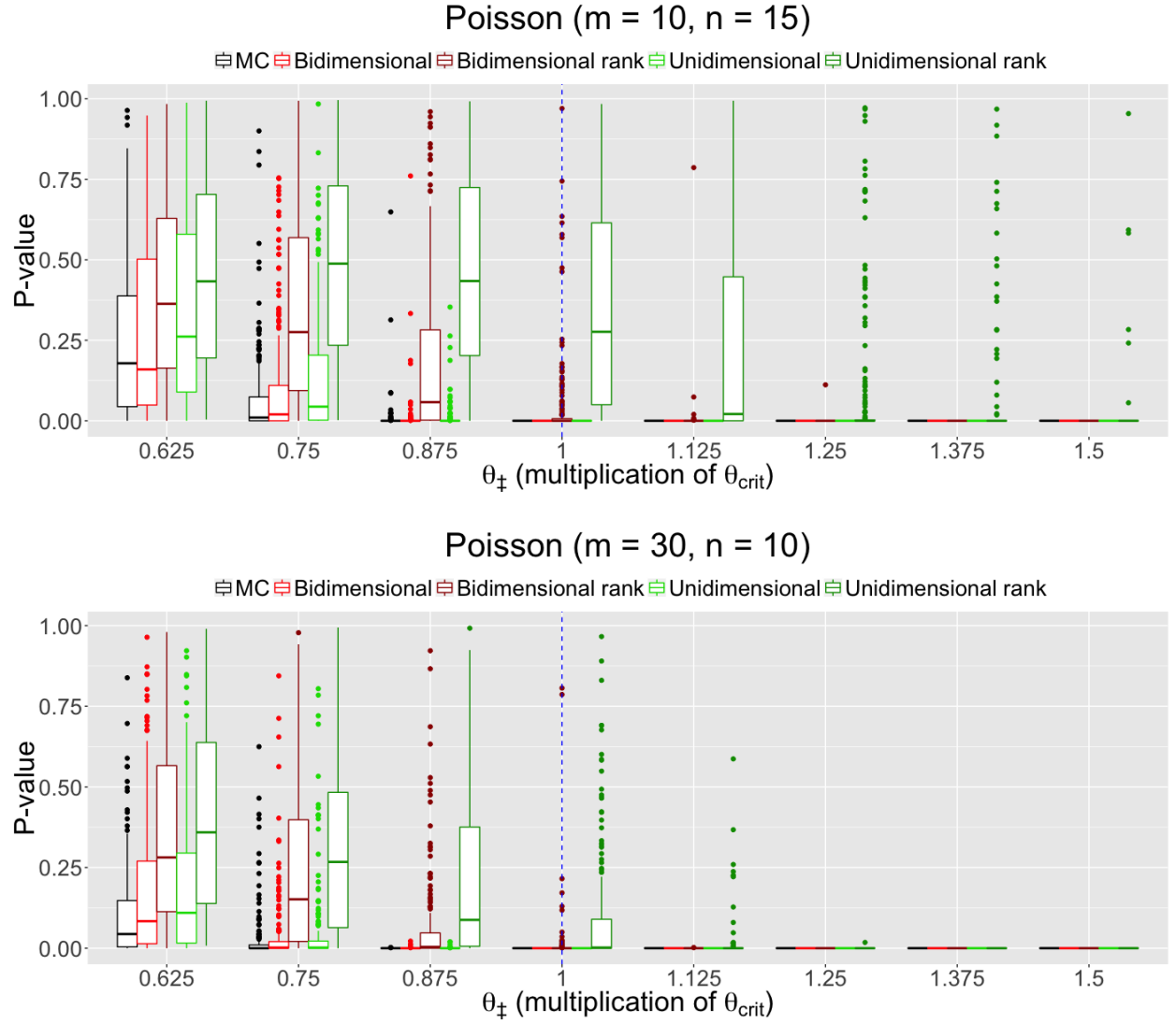


Figure 2: P-values of various forms of scan tests in the Poisson model

6 Proofs

6.1 Preliminaries

We start with some preliminary results that already appear, one or another, in our previous work (Arias-Castro et al., 2015). First, for any one-parameter exponential family $(f_\theta : \theta \in \Theta)$ with a standardized base distribution ν , as we consider to be here,

$$\mathbb{E}_\theta(X) \geq \theta, \quad \forall \theta \in \Theta. \quad (18)$$

Next, in the same context, if $\sup \Theta > 0$ (which we assume throughout), then f_θ has a sub-exponential right tail, which is uniform in $\theta \in \bar{\theta}$ if $\bar{\theta} \in \Theta$. In particular, there is $\bar{\gamma}$ that depends on $\bar{\theta} > 0$ such that, if X_1, \dots, X_k are independent, with $X_j \sim f_{\theta_j}$ with $\theta_j \leq \bar{\theta}$, then

$$\max_{j \in [k]} X_j \leq \bar{\gamma} \log k, \quad \text{with probability tending 1 as } k \rightarrow \infty. \quad (19)$$

By symmetry, if $\inf \Theta < 0$ (which we assume in the case of unidimensional permutations), the same is true on the left. In particular, ν itself (corresponding to $\theta = 0$) has a sub-exponential left tail in this case, and this is all that will be used below. In particular, there is a constant $\gamma_0 > 0$ such that, if X_1, \dots, X_k are IID ν , then

$$\min_{j \in [k]} X_j \geq -\gamma_0 \log k, \quad \text{with probability tending 1 as } k \rightarrow \infty. \quad (20)$$

6.2 Proof of Theorem 2

The proof arguments are parallel to those of Arias-Castro et al. (2015), derived in the context of more structured settings. For that reason, we only detail the proof of Theorem 2 in the case of unidimensional permutations, which, compared to bidimensional permutations, is a bit more different from the setting considered in (Arias-Castro et al., 2015) and requires additional arguments. Therefore, in what follows, we take $\Pi = \Pi_1$. Recall that in this case we assume in addition that $\varphi(\theta) < \infty$ for some $\theta < 0$. This implies that ν has sub-exponential tails

Case (i) We first focus on the condition where ν has support bounded from above and let b_0 denote such an upper bound. (Necessarily, $b_0 > 0$.) Thus, regardless of the θ_{ij} 's,

$$\mathbb{P}(\max_{i,j} X_{ij} \leq b_0) = 1. \quad (21)$$

The permutation scan test has limiting power 1 if and only if $\mathbb{P}(\mathfrak{P}(\mathbf{X}) \leq \alpha) \rightarrow 1$ under the alternative. We show that by proving the stronger claim that $\mathfrak{P}(\mathbf{X}) \rightarrow 0$ in probability under the alternative.

We first work conditional on $\mathbf{X} = \mathbf{x}$, where $\mathbf{x} = (x_{ij})$ denotes a realization of $\mathbf{X} = (X_{ij})$. We may equivalently center the rows of \mathbf{X} before scanning, and the resulting test remains unchanged. Therefore, we may assume that all the rows of \mathbf{x} sum to 0. Let $\zeta = \text{SCAN}(\mathbf{x})$ for short. We have

$$\mathfrak{P}(\mathbf{x}) = \mathbb{P}(\text{SCAN}(\mathbf{x}_\pi) \geq \zeta), \quad (22)$$

where the randomness comes solely from π , uniformly drawn from Π . Using the union bound, we get

$$\mathfrak{P}(\mathbf{x}) \leq |\mathbb{S}_{m,n}| \max_{\mathcal{S} \in \mathbb{S}_{m,n}} \mathbb{P}\left(\sum_{(i,j) \in \mathcal{S}} x_{\pi(i,j)} \geq \zeta\right). \quad (23)$$

For each $i \in [N]$, let $(A_{ij} : j \in [n])$ be a sample from $(x_{ij} : j \in [N])$ *without* replacement and let $A_i = \sum_{j \in [n]} A_{ij}$. Note that A_1, \dots, A_M are independent and, for $\mathcal{S} = \mathcal{I} \times \mathcal{J}$, we have

$$\sum_{(i,j) \in \mathcal{S}} x_{\pi(i,j)} \sim \sum_{i \in \mathcal{I}} A_i. \quad (24)$$

Fix $\mathcal{I} \subset [M]$ of size m . Using Markov's inequality and the independence of the A_i 's, we get

$$\mathbb{P}(\sum_{i \in \mathcal{I}} A_i \geq \zeta) \leq e^{-c\zeta} \prod_{i \in \mathcal{I}} \phi_i(c), \quad (25)$$

where ϕ_i is the moment generating function of A_i . The key is (Hoeffding, 1963, Th 4), which implies that $\phi_i \leq \psi_i$, where ψ_i is the moment generating function of B_i , where $B_i = \sum_{j \in [n]} B_{ij}$ and $(B_{ij} : j \in [n])$ is a sample from $(x_{ij} : j \in [N])$ *with* replacement, meaning that these are IID random variables uniformly distributed in $(x_{ij} : j \in [N])$. By (21), we have $B_{ij} \leq b_0$ with probability one, and the usual arguments leading to the (one-sided) Bernstein's inequality yield the usual bound

$$\psi_i(c) \leq \exp \left[\frac{nc^2 \sigma_i^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2 / 2} \right], \quad (26)$$

where σ_i^2 is the variance of B_{i1} , meaning, $\sigma_i^2 = \frac{1}{N} \sum_{j \in [N]} (x_{ij} - \bar{x}_i)^2$, with $\bar{x}_i = \frac{1}{N} \sum_{j \in [N]} x_{ij}$ being the mean. Letting $\sigma^2 = \max_{i \in [M]} \sigma_i^2$, we derive

$$e^{-c\zeta} \prod_{i \in \mathcal{I}} \phi_i(c) \leq e^{-c\zeta} \prod_{i \in \mathcal{I}} \exp \left[\frac{nc^2 \sigma_i^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2 / 2} \right] \leq e^{-c\zeta} \exp \left[\frac{mnc^2 \sigma^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2 / 2} \right], \quad (27)$$

the latter being the usual bound that leads to Bernstein's inequality. The same optimization over c yields

$$\mathbb{P}(\sum_{i \in \mathcal{I}} A_i \geq \zeta) \leq \exp \left[-\frac{\zeta^2}{2mn\sigma^2 + \frac{2}{3}b_0\zeta} \right]. \quad (28)$$

We now emphasize the dependency of ζ and σ^2 on \mathbf{x} by adding \mathbf{x} as a subscript. Noting that this bound is independent of \mathcal{I} (of size m), we get

$$\mathfrak{P}(\mathbf{x}) \leq |\mathbb{S}_{m,n}| \exp \left[-\frac{\zeta_{\mathbf{x}}^2}{2mn\sigma_{\mathbf{x}}^2 + \frac{2}{3}b_0\zeta_{\mathbf{x}}} \right]. \quad (29)$$

We now free \mathbf{X} and bound $\zeta_{\mathbf{X}}$ from below, and $\sigma_{\mathbf{X}}^2$ from above. When doing so, we need to take into account that we assumed the rows summed to 0. When this is no longer the case, $\zeta_{\mathbf{X}}$ denotes the scan of \mathbf{X} after centering all the rows. Let \bar{X}_i denote the mean of row i . By definition of the scan in (3),

$$\zeta_{\mathbf{X}} \geq \zeta_{\text{true}} := \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \in \mathcal{J}_{\text{true}}} (X_{ij} - \bar{X}_i) = (1 - \frac{n}{N}) \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \in \mathcal{J}_{\text{true}}} X_{ij} - \frac{n}{N} \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \notin \mathcal{J}_{\text{true}}} X_{ij}. \quad (30)$$

For the expectation, by (8) and (18), we have

$$\mathbb{E}(\zeta_{\text{true}}) \geq (1 - \frac{n}{N}) \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \in \mathcal{J}_{\text{true}}} \theta_{ij} \geq (1 - o(1))mn\theta_{\dagger}. \quad (31)$$

For the variance, we have $\text{Var}(X_{ij}) = 1$ when $(i, j) \notin \mathcal{S}_{\text{true}}$ (since ν has variance 1) and $\text{Var}(X_{ij}) \leq \mathbb{E}(X_{ij}^2) \leq b_0^2$ always. Using this, we derive

$$\text{Var}(\zeta_{\text{true}}) \leq mnb_0^2 + (\frac{n}{N})^2 mN = mn(b_0^2 + \frac{n}{N}) = O(mn). \quad (32)$$

Because of (8) and (10), $\mathbb{E}(\zeta_{\text{true}}) \gg \sqrt{\text{Var}(\zeta_{\text{true}})}$, and thus by Chebyshev's inequality,

$$\zeta_{\text{true}} = (1 + o_P(1)) \mathbb{E}(\zeta_{\text{true}}) \geq (1 + o_P(1)) mn\theta_{\dagger}^2. \quad (33)$$

We now bound $\sigma_{\mathbf{X}}^2$. For $i \in \mathcal{I}_{\text{true}}$, we have

$$\sigma_i^2(\mathbf{X}) \leq \frac{1}{N} \sum_{j \in [N]} X_{ij}^2 = \frac{1}{N} \sum_{j \in \mathcal{J}_{\text{true}}} X_{ij}^2 + \frac{1}{N} \sum_{j \notin \mathcal{J}_{\text{true}}} X_{ij}^2 \leq \frac{nb_0^2}{N} + \frac{1}{N} \sum_{j \notin \mathcal{J}_{\text{true}}} X_{ij}^2. \quad (34)$$

For $i \notin \mathcal{I}_{\text{true}}$,

$$\sigma_i^2(\mathbf{X}) \leq \frac{1}{N} \sum_{j \in [N]} X_{ij}^2. \quad (35)$$

Therefore

$$\sigma_{\mathbf{X}}^2 \stackrel{\text{sto}}{\leq} 1 + o(1) + \max_{i \in [M]} \frac{1}{N} \sum_{j \in [N]} T_{ij}, \quad (36)$$

where $(T_{ij} : (i, j) \in [M] \times [N])$ are IID with distribution that of $X^2 - 1$ when $X \sim \nu$. Note that $\mathbb{E}(T_{ij}) = 0$ since ν has variance 1 and

$$\max_{i,j} T_{ij} \leq \bar{t} := b_0^2 \vee (\gamma_0 \log(MN))^2, \quad (37)$$

by (21) and when the following event holds

$$\mathcal{A} := \left\{ \min_{i,j} X_{ij} \geq -\gamma_0 \log(MN) \right\}, \quad (38)$$

which by (20) happens with probability tending to 1. Let $\mathbb{P}_{\mathcal{A}}$ be the probability conditional on \mathcal{A} and $\mathbb{E}_{\mathcal{A}}$ the corresponding expectation. Let $\mu_{\mathcal{A}} = \mathbb{E}_{\mathcal{A}}(T_{ij})$ and $\tau_{\mathcal{A}}^2 = \text{Var}_{\mathcal{A}}(T_{ij}) < \infty$, because ν has finite fourth moment. By Bernstein's inequality, for any $c > \mu_{\mathcal{A}}$,

$$\mathbb{P}_{\mathcal{A}} \left(\frac{1}{N} \sum_{j \in [N]} T_{ij} > c \right) \leq \exp \left[- \frac{N(c - \mu_{\mathcal{A}})^2}{2\tau_{\mathcal{A}}^2 + \frac{2}{3}\bar{t}c} \right]. \quad (39)$$

Then using a union bound

$$\mathbb{P}_{\mathcal{A}} \left(\max_{i \in [M]} \frac{1}{N} \sum_{j \in [N]} T_{ij} > c \right) \leq M \exp \left[- \frac{N(c - \mu_{\mathcal{A}})^2}{2\tau_{\mathcal{A}}^2 + \frac{2}{3}\bar{t}c} \right]. \quad (40)$$

Taking logs, noting that $\mu_{\mathcal{A}} \rightarrow 0$ and $\tau_{\mathcal{A}}^2 \rightarrow \tau^2 := \text{Var}(T_{ij})$, as well as $\bar{t} = O(\log(MN))$, and using (8) and (13), we see that the RHS tends to 0 for any $c > 0$ fixed. Therefore $\max_{i \in [M]} \frac{1}{N} \sum_{j \in [N]} T_{ij} = o_P(1)$ conditional on \mathcal{A} , and since $\mathbb{P}(\mathcal{A}) \rightarrow 1$, also unconditionally. Coming back to (36), we conclude that

$$\sigma_{\mathbf{X}}^2 = 1 + o_P(1). \quad (41)$$

The upper bound on $\zeta_{\mathbf{X}}$ and the lower bound on $\sigma_{\mathbf{X}}^2$, combined, imply by monotonicity that

$$\frac{\zeta_{\mathbf{X}}^2}{2mn\sigma_{\mathbf{X}}^2 + \frac{2}{3}b_0\zeta_{\mathbf{X}}} \geq (1 + o_P(1)) \frac{mn\theta_{\dagger}^2}{2 + \frac{2}{3}b_0\theta_{\dagger}^2}. \quad (42)$$

We also have $|\mathbb{S}_{m,n}| = \binom{M}{m} \binom{N}{n}$, so that

$$\log |\mathbb{S}_{m,n}| = \log \binom{M}{m} + \log \binom{N}{n} \leq (1 + o(1)) \Lambda, \quad (43)$$

with

$$\Lambda := \left\lceil m \log \frac{M}{m} + n \log \frac{N}{n} \right\rceil, \quad (44)$$

where in the last inequality we used (8) and the fact that $\log \binom{K}{k} \leq k \log(K/k) + k$ for all integers $1 \leq k \leq K$.

Coming back to (29) and collecting all the bounds in between, we find that

$$\log \mathfrak{P}(\mathbf{X}) \leq (1 + o(1))\Lambda - (1 + o_P(1)) \frac{mn\theta_{\dagger}^2}{2 + \frac{2}{3}b_0\theta_{\dagger}}. \quad (45)$$

Under (10), there is $\varepsilon > 0$ such that, eventually,

$$\theta_{\dagger} \geq (1 + \varepsilon)\sqrt{2\Lambda/(mn)}. \quad (46)$$

When that's the case, we get

$$\log \mathfrak{P}(\mathbf{X}) \leq (1 + o(1))\Lambda - (1 + o_P(1)) \frac{(1 + 2\varepsilon)\Lambda}{1 + \frac{1}{3}b_0(1 + \varepsilon)\sqrt{2\Lambda/(mn)}}. \quad (47)$$

Noting that $\Lambda/(mn) = o(1)$ and $\Lambda \rightarrow \infty$ under (8), we get

$$\log \mathfrak{P}(\mathbf{X}) \leq -(1 + o_P(1))2\varepsilon\Lambda \rightarrow -\infty, \quad (48)$$

which is what we needed to prove.

Case (ii) We now consider the case where $\theta_{ij} \leq \bar{\theta}$ for all $(i, j) \in [M] \times [N]$ for some $\bar{\theta} < \theta_*$. Although (21) may not hold for any b_0 , we redefine $b_0 = \bar{\gamma} \log(MN)$, where $\bar{\gamma}$ depends on $\bar{\theta}$, and condition on the event

$$\mathcal{B} := \left\{ \max_{i,j} X_{ij} \leq b_0 \right\}, \quad (49)$$

which holds with probability tending to 1 by (19). The bound (29) holds unchanged (assuming that $\max_{i,j} x_{ij} \leq b_0$). What is different is how $\zeta_{\mathbf{X}}$ and $\sigma_{\mathbf{X}}^2$ are handled, now that we conditioned on \mathcal{B} . Let $\mathbb{P}_{\mathcal{B}}$ and $\mathbb{E}_{\mathcal{B}}$ denote the probability and expectation conditional on \mathcal{B} .

We have

$$\mathbb{E}_{\mathcal{B}}(\zeta_{\text{true}}) \geq (1 - \frac{n}{N}) \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \in \mathcal{J}_{\text{true}}} \mathbb{E}_{\mathcal{B}}(X_{ij}) - \frac{n}{N} \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \notin \mathcal{J}_{\text{true}}} \mathbb{E}_{\mathcal{B}}(X_{ij}) \quad (50)$$

$$\geq (1 - \frac{n}{N}) \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \in \mathcal{J}_{\text{true}}} \mathbb{E}(X_{ij} | X_{ij} \leq b_0) - \frac{n}{N} \sum_{i \in \mathcal{I}_{\text{true}}} \sum_{j \notin \mathcal{J}_{\text{true}}} \mathbb{E}(X_{ij} | X_{ij} \leq b_0) \quad (51)$$

$$\geq (1 + o(1))mn\theta_{\dagger}. \quad (52)$$

In the last inequality, for $j \notin \mathcal{J}_{\text{true}}$ we used the fact that $\mathbb{E}(X_{ij}) = 0$, which implies that $\mathbb{E}_{\mathcal{B}}(X_{ij}) \leq 0$ in that case. And for $j \in \mathcal{J}_{\text{true}}$ we used the fact that $\mathbb{E}(X_{ij} | X_{ij} \leq b_0) \rightarrow \theta_{ij} \geq \theta_{\dagger}$ combined with a Cèsaro-type argument. On the other hand, in a similar way, we also have

$$\text{Var}_{\mathcal{B}}(\zeta_{\text{true}}) = O(mnb_0^2) = O(mn \log^2(MN)). \quad (53)$$

So we still have $\mathbb{E}_{\mathcal{B}}(\zeta_{\text{true}}) \gg \sqrt{\text{Var}_{\mathcal{B}}(\zeta_{\text{true}})}$, by (8) and (10), and in addition (13). In particular, (33) holds under \mathcal{B} . In very much the same way, one can verify that the same is true of (41).

From there we get to (47) in exactly the same way, conditional on \mathcal{B} , and then unconditionally since $\mathbb{P}(\mathcal{B}) \rightarrow 1$. Then, to conclude, we only need to check that $b_0\sqrt{\Lambda/(mn)} = o(1)$, which is the case by (13).

Acknowledgements

This work was partially supported by a grant from the US Office of Naval Research (N00014-13-1-0257) and a grant from the US National Science Foundation (DMS 1223137).

References

- Arias-Castro, E., R. M. Castro, E. Tánzos, and M. Wang (2015). Distribution-free detection of structured anomalies: Permutation and rank-based scans. *arXiv preprint arXiv:1508.03002*.
- Barry, W. T., A. B. Nobel, and F. A. Wright (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21(9), 1943–1949.
- Butucea, C. and Y. I. Ingster (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* 19(5B), 2652–2688.
- Cai, T. T., T. Liang, and A. Rakhlin (2015). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*.
- Chen, Y. and J. Xu (2014). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*.
- Cheng, Y. and G. M. Church (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103. AAAI Press.
- Hájek, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics* 33(3), 1124–1147.
- Hájek, J. and Z. Sidak (1967). *Theory of rank tests*. Academic Press, Academia Publishing House of the Czechoslovak Acad.
- Hastie, T., R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown (2000). ‘gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1(2), 1–21.
- Hemerik, J. and J. Goeman (2014). Exact testing with random permutations. *arXiv preprint arXiv:1411.7565*.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58, 13–30.
- Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer.
- Ma, Z. and Y. Wu (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics* 43(3), 1089–1116.
- Madeira, S. C. and A. L. Oliveira (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 1(1), 24–45.
- Shabalín, A. A., V. J. Weigman, C. M. Perou, and A. B. Nobel (2009). Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics* 3(3), 985–1012.